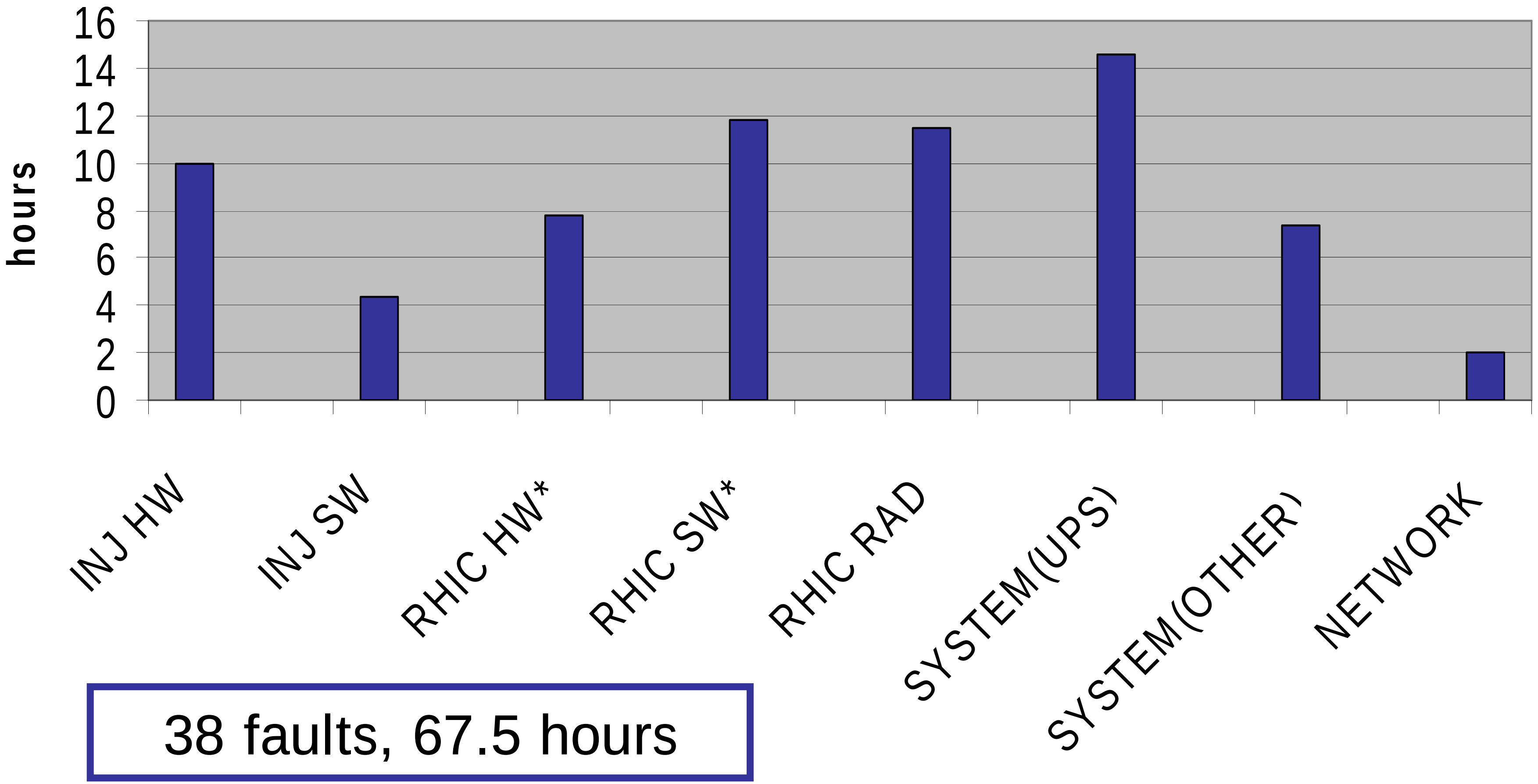
Controls Software Availability & Reliability

- 2006 Controls Downtime Statistics
- Responses to 2006 Problem Areas
- Long Term Strategies
- Summary

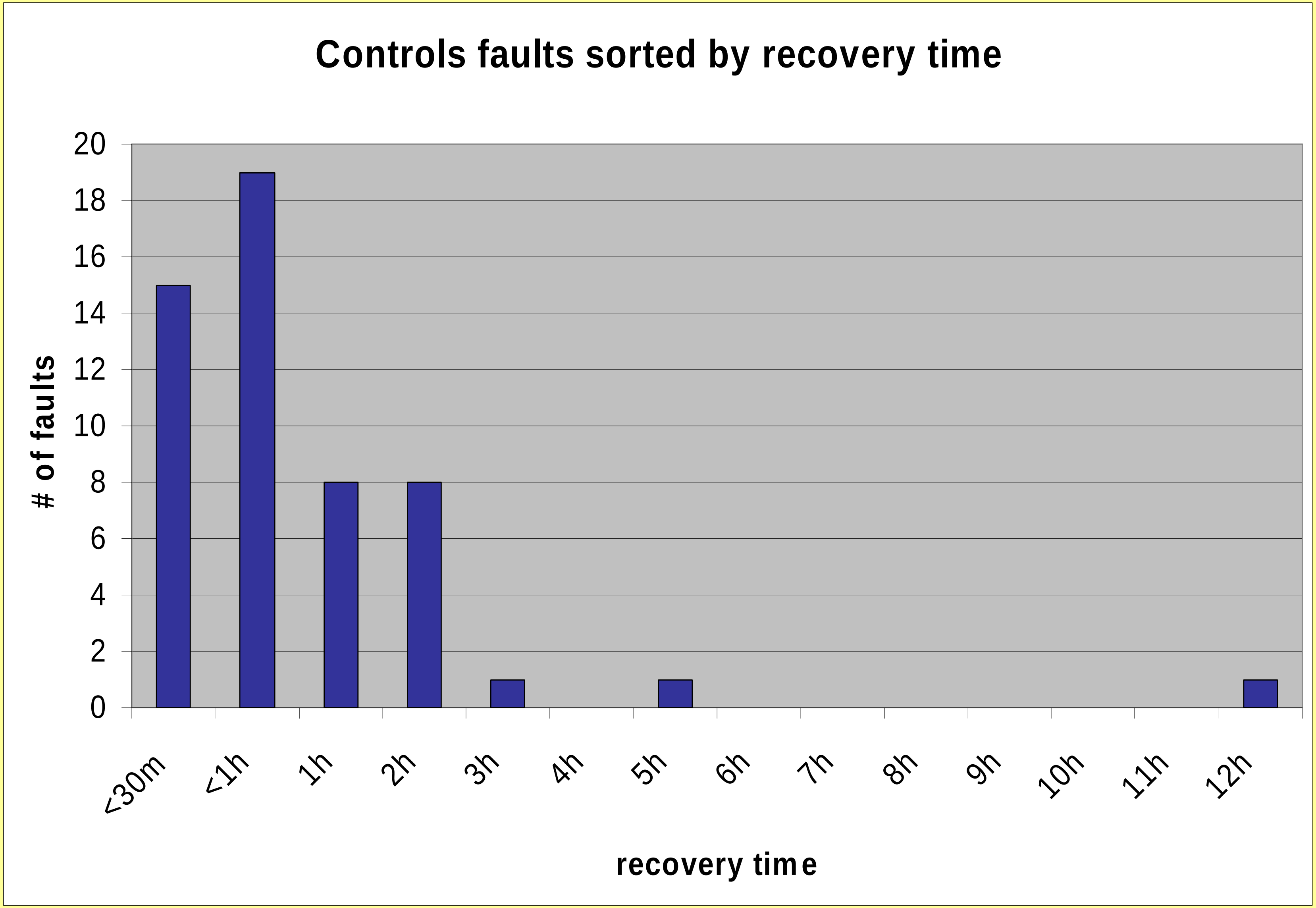
John T. Morris

July 8, 2006

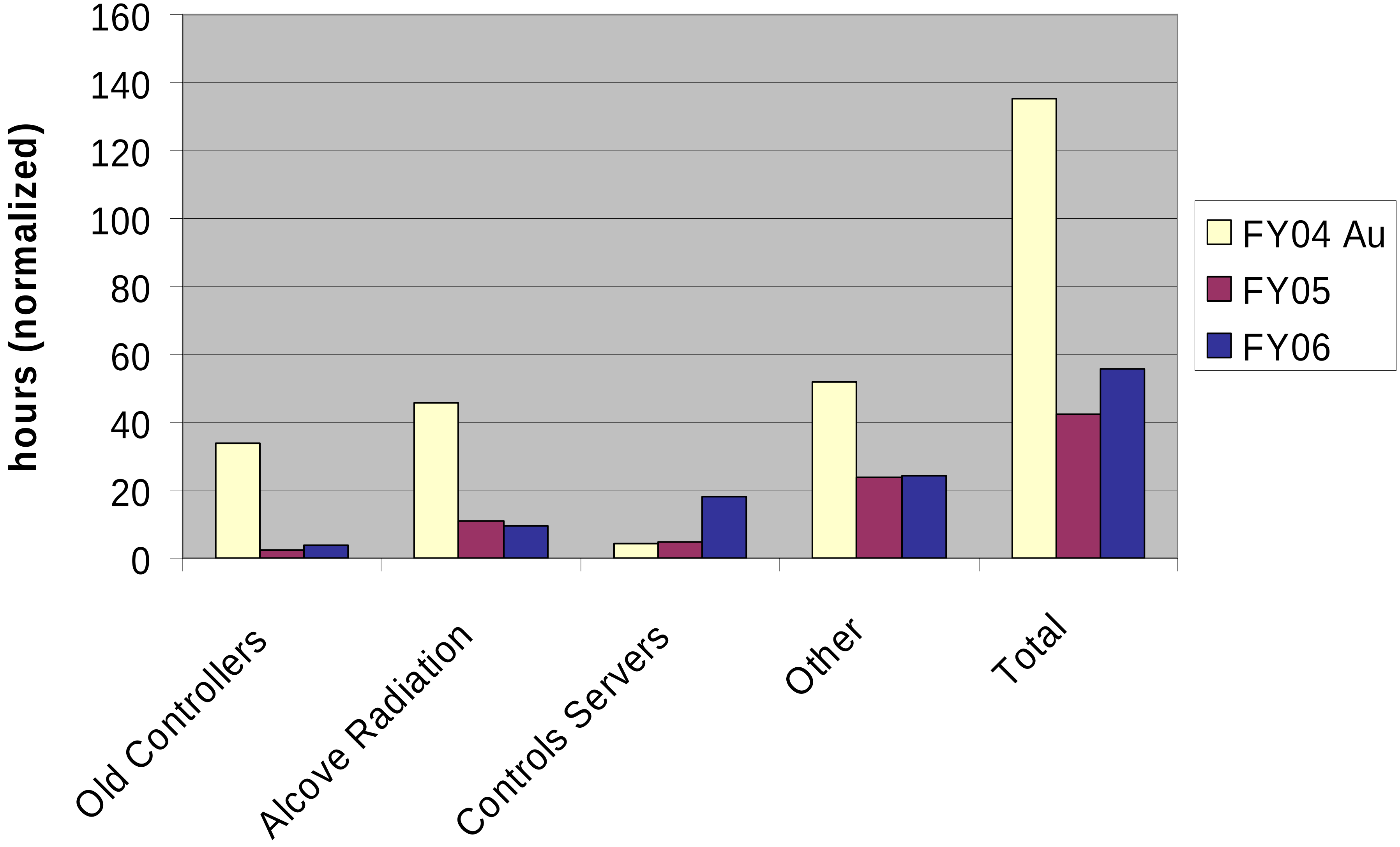
Major Categories of Controls Downtime for RHIC
1/15/06 to 6/26/06



38 faults, 67.5 hours


Controls faults sorted by recovery time
# of faults
20
18
16
14
12
10
8
6
4
2
0
<30m
<1h
1h
2h
3h
4h
5h
6h
7h
8h
9h
10h
11h
12h
recovery time

Controls Failure Hours Normalized to 19 Week Period

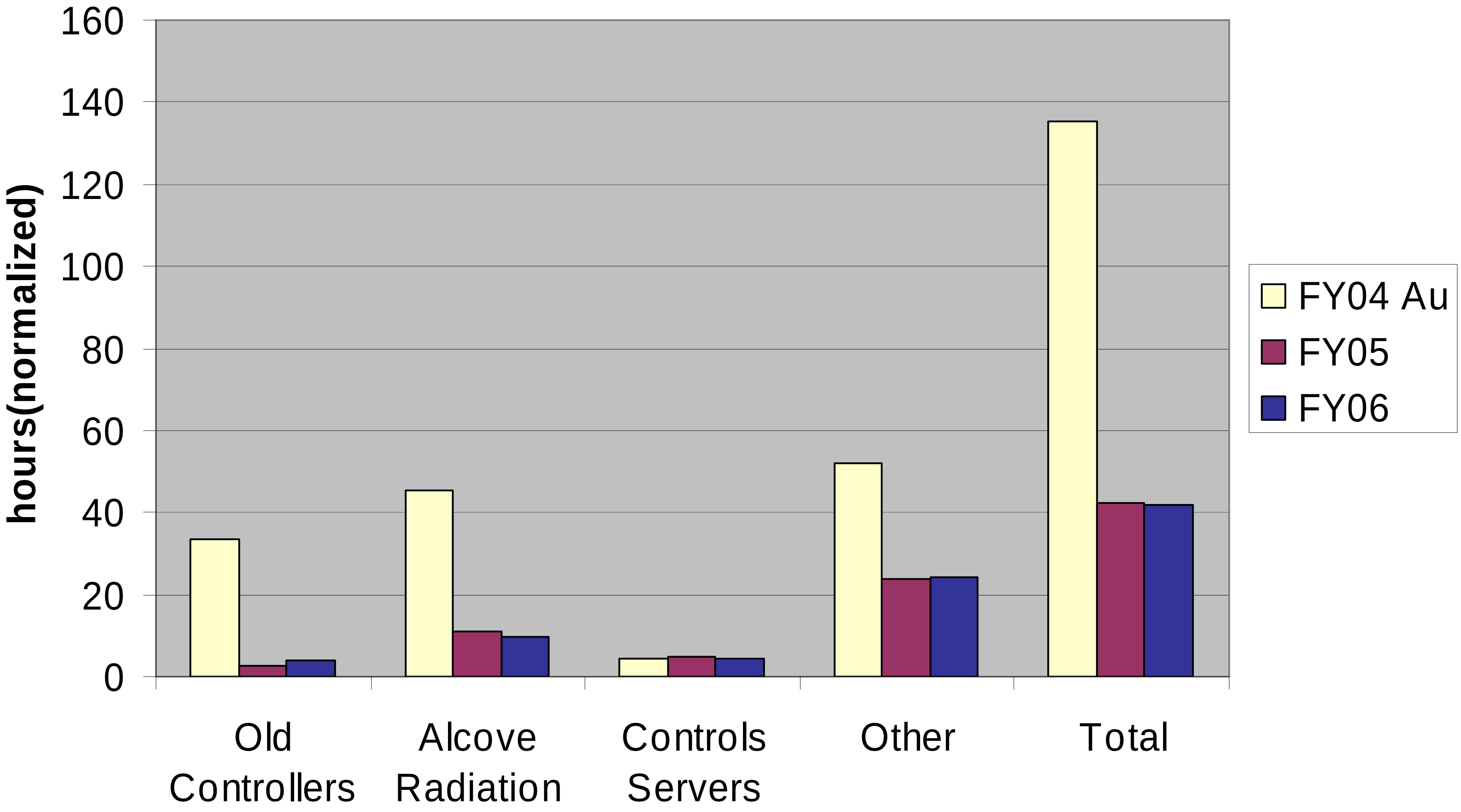


Why the Increase in Failure Hours for Controls Servers?

- Power interruptions to computer room
- Operations RAID controller failover not working properly –> file system corruption

Note that there were very few problems due to Linux Red Hat OS instability (a major problem in 2005)

What if 2006 UPS & RAID failures had not occurred...



Responses to major failures

- Power interruption/ Op RAID failure (16.7h)
  - Computer room UPS upgrade (7/17)
  - Disaster recovery planning/testing (all summer)
    - Server reconfiguration
    - Improved recovery tools
    - Test and document procedures
    - Testing will be disruptive – primarily evening hours
  - Op RAID troubleshooting (all summer)
    - Diagnostic dumps / fibre channel analyzer
    - System will be taken offline & stress tested

Responses to major failures

- BLAM “double count” (2.5h)
  - Data correlation problem found & fixed
- AgsOrbitControl restore problems (2.2h)
  - bug found and fixed
  - For 2007: review & revise AgsOrbitControl function delivery & archive management

Responses to major failures

- Cryo data delivery failures (2.1h)
  - Misassigned data: controls bug fixed
  - Delivery interruptions: workaround for Linux bug, redundant server added, comm tuning
  - For 2007: dedicated host for cryo servers
  - *Can we reduce vulnerability at cryo end?*
- Quench detect FEC failures (2.0h)
  - Vxworks investigation, network reconfiguration
  - Summer ‘06: reproduce in lab & add protection

Other problems

- Recover from rad upsets
  - *Chassis ps replacement (8h/4 events)*
  - Other recovery (3h/3 events)
- Linac file locking problem (3.4h)
  - Triggered by troubleshooting after UPS recovery
  - Sys admin lessons learned
- All other software/system (8.3h/17 events)
  - RhicInjection app problems (.9h/2 events)
  - Polarimeter control problems (.8h/2 events)
  - Server reboots (.9h/2 events)

Long term strategies :
How to avoid SW/system downtime

- Communicate with operations re releases
- Test well before release
  - Only 3 faults in 2006 associated with new SW (3.6h)
  - No faults due to frivolous or untested releases
- Design with errors & unusual conditions in mind
  - 20/20 hindsight – Some SW could have been designed a priori to avoid some 2006 problems
  - When is technical review of SW appropriate?
  - Recognize tradeoff – sacrifice productivity, delay forward progress. Target critical systems

Long term strategies : How to avoid SW/system downtime

- Have fallback SW versions readily available
- Provide prompt support to solve problems
- Give operations troubleshooting tools
  - Effective: Recovery time for most controls faults is < 1 hour

---

- Organize teams to attack difficult problems
  - Effective in 2006. Could have been faster in some cases (e.g. BLAM)

Long term strategies : How to avoid SW/system downtime

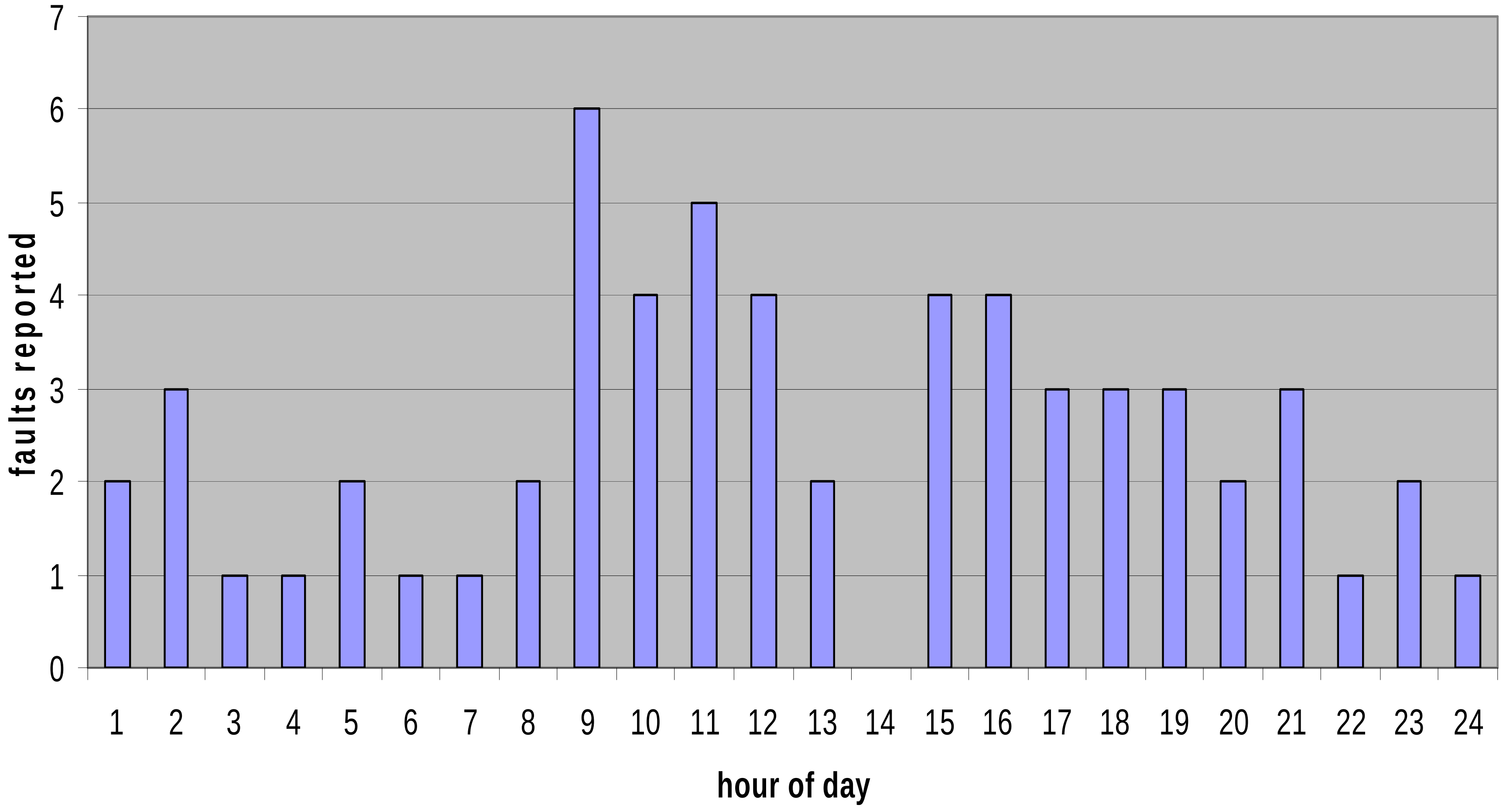
- Aggressively implement solutions to downtime vulnerabilities
  - Has to compete for resources with new systems and upgrades (and win!)
- Pay attention to vulnerabilities, not just failure history (try to pick the right ones)
  - Legacy systems
  - Rad upsets of wfgs
- Provide redundancy for critical components
  - Switch to fallback automatic or “easy”/documented

Summary

- Controls SW/system downtime dominated by UPS and RAID problems in 2006
- Work underway to address UPS/RAID and general disaster recovery issues
- Other problem areas already addressed for the most part – some continuing work
- Long term strategies should reduce downtime
- Avoiding ctls downtime can not be sole concern
    - “change nothing” is not an option
    - New development can help overall machine reliability/availability

The end

Controls Faults Started During Each Hour of Day



Faults reported between 9 & 11am.

- 01-23 09:30 dh158,ps bus ctlr & fan replaced
- 02-14 09:34 Blue link pulled by 2b-ps1Yellow link pulled by 8b-ps1 *(still really 6b UPS)*
- 02-14 10:52 5A-SW13 not working alarm
- 02-14 10:59 UPS - cfe-6b-ps1 no heartbeat
- 03-08 09:35 lin84 crash/reboot
- 04-03 10:00 Sequencer locked up. Can't ramp down. Polarimeter sequence hanging up
- 04-04 09:35 UPS - Restoring ctrls to MCR
- 04-23 09:30 Collimators stuck in the in position
- 05-08 10:26 CryoWrite Server-- root cause

Beyond down time – How can we facilitate more reliable and reproducible accelerator operation?

- Better diagnostics in injectors
  - Injector snapshots & ‘agscompare’
  - ‘pswatch’ alarms for set/ref/current mismatch in injector functions
  - Replacement of old equipment
  - Post mortem for more Injector systems (e.g. AGS main magnet)
- More reliable & transparent control of AGS orbit correctors
- Improved diagnostics for polarimeter systems
- Logging all info for ATR shots